hw12

Zhuoyu Jiang

12/6/2022

Question 1

a.

Since the time period is 8 am to 9 am, t=1 Count 225 car crossing the bridge, n=225

```
alpha<-0.05
n<-225
t<-1
lambda.hat<-n/t
lambda.hat+qnorm(p=c(alpha/2,1-alpha/2))*sqrt(lambda.hat/n)
```

```
## [1] 223.04 226.96
```

Thus, with 95% confidence interval, the confident interval for the true mean car for this time period is between 223.04 and 226.96.

b.

The observation and prediction results in one time period are difficult to be extended to other time periods, and more researches should be taken, which increases the research cost. The variance (which is same as λ) of the observation data would be very large, and the true confidence interval would be even wider.

Question 2

a.

Time of day: Normally, the rush hour happens twice every weekday: once in the morning and once in the afternoon or evening, the times during which the most people commute. During the rush hour, the number of cars stopped at the intersection would increase.

Day of the week: People tend to have long travel on weekends, the number of cars stopped at the intersection would increase on weekdens.

Whether school in session: Normally, during school in session, cars share the roads once more with buses and students, and more congestion on the roads. Thus, the number of cars stopped at the intersection would increase.

When designing the experiment, it can be considered to do eight experiments: the rush hour of weekends while school in session, the valley hour of weekends while school in session, the rush hour of one weekday while school in session, the valley hour of one weekday while school in session, the rush hour of weekends while school not in session, the valley hour of weekends while school not in session, the rush hour of one weekday while school not in session, the valley hour of one weekday while school not in session. and the result would be representative.

b.

This is not a true Poisson process because the arrivals are not independent of one another. Even for bus systems that do not run on time, whether or not one bus is late affects the arrival time of the next bus. Thus, we should consider: 1. The number of cars that arrive at the intersection can be counted. 2. The arrival of one car does not affect the arrival of another car. 3. We can easily collect data on the average number of cars that arrive the intersection. 4. Two cars cannot occur at exactly the same instant in time.

c.

$$P(k \text{ events in time period}) = e^{-\lambda * \frac{\lambda^k}{k!}}$$

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
Stoplight <- read_csv("~/Downloads/Stoplight.csv")
```

```
## Rows: 40 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Observation, vehicles
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
k<-9:15
lambda<-mean(Stoplight$vehicles)
1-ppois(8,lambda = lambda)
```

```
## [1] 0.01786982
```

During 3:25pm to 4:05 pm on a non-holiday weekday, the probability of more than 9 vehicles showing in one stoplight cycle is 0.01786982.

d.

```
p<-0.01786982
n<-60
1-pbinom(0, size=n, prob=p)
```

```
## [1] 0.6610439
```

The probability that the fire station's driveway is at least partially blocked at least once over 60 cycles of the light is 0.6610439.

Question 3

$$\mu_i = exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip})$$

Adding the function to the equation in 4.1.1

$$L(\beta; y, x) = \prod_{i=0}^{n} \frac{e^{-exp(x_i\beta)} exp(x_i\beta)^{y_i}}{y_i!}$$

Deriving it,

$$l(\beta) = \sum_{i=0}^{n} y_i x_i \beta - \sum_{i=0}^{n} exp(x_i\beta) - \sum_{i=0}^{n} log(y_i!)$$

Question 4

a.

```
dt <- read_csv("~/Desktop/dt.csv")
```

```
## Rows: 4406 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (8): ofp, hosp, numchron, gender, school, privins, health_excellent, hea...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
model.fit.dt <- glm(ofp~hosp+numchron+gender+school+privins+health_excellent+health_poor, family="poiss
summary(model.fit.dt)
```

```
##
## Call:
## glm(formula = ofp ~ hosp + numchron + gender + school + privins +
##     health_excellent + health_poor, family = "poisson", data = dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4055  -1.9962  -0.6737   0.7049  16.3620
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.028874   0.023785  43.258   <2e-16 ***
## hosp              0.164797   0.005997  27.478   <2e-16 ***
## numchron          0.146639   0.004580  32.020   <2e-16 ***
```

```
## gender           -0.112320   0.012945  -8.677   <2e-16 ***
## school            0.026143   0.001843  14.182   <2e-16 ***
## privins           0.201687   0.016860  11.963   <2e-16 ***
## health_excellent -0.361993   0.030304 -11.945   <2e-16 ***
## health_poor       0.248307   0.017845  13.915   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 23168  on 4398  degrees of freedom
## AIC: 35959
##
## Number of Fisher Scoring iterations: 5
```

The Poisson regression model is

$exp(ofp) = 1.028874 + 0.164797 * hosp + 0.146639 * numchron - 0.112320 * gender + 0.026143 * school + 0.201687 * privins - 0.36199$

b.

The coefficient of hospital stay is 0.164797, and $e^{0.164797} > 1$. The expected count $\mu = E(office\ visit\ for\ a\ person)$ is 0.164797 times larger than when hospital stay is 0.

The coefficient of number of chronic conditions is 0.146639, and $e^{0.146639} > 1$. The expected count $\mu = E(office\ visit\ for\ a\ person)$ is exp(0.14663) times larger than when number of chronic conditions is 0.

The coefficient of gender is -0.112320, and $e^{-0.112320} < 1$. The expected count $\mu = E(office\ visit\ for\ a\ person)$ is exp(-0.112320) times smaller than when gender is 0.

The coefficient of school is 0.026143, and $e^{0.026143} > 1$. The expected count $\mu = E(office\ visit\ for\ a\ person)$ is exp(0.026143) times larger than when school is 0.

The coefficient of private insurance is 0.201687, and $e^{0.201687} > 1$. The expected count $\mu = E(office\ visit\ for\ a\ person)$ is exp(0.201687) times larger than when private insurance is 0.

The coefficient of health_excellent is -0.361993, and $e^{-0.361993} < 1$. The expected count $\mu = E(office\ visit\ for\ a\ person)$ is exp(-0.361993) times smaller than when health_excellent is 0.

The coefficient of health_poor is 0.248307, and $e^{0.248307} < 1$. The expected count $\mu = E(office\ visit\ for\ a\ person)$ is exp(0.248307) times larger than when health_poor is 0.

c.

```
data1=dt[which(dt$ofp==0),]
data1
```

```
## # A tibble: 683 x 8
##      ofp  hosp numchron gender school privins health_excellent health_poor
##    <dbl> <dbl>    <dbl>  <dbl>  <dbl>   <dbl>            <dbl>       <dbl>
##  1     0     0        0      0      8       1                0           0
##  2     0     0        1      1      8       1                0           0
##  3     0     0        1      1      8       0                0           0
```

```
##  4     0     0        0      1      9       1                0           0
##  5     0     0        0      1     13       1                0           0
##  6     0     0        0      0      0       0                0           0
##  7     0     1        0      1     18       1                0           0
##  8     0     0        0      0      9       1                0           0
##  9     0     0        1      1      8       1                0           0
## 10     0     0        1      0      9       1                0           0
## # ... with 673 more rows
```

The reason why so many zero-visit counts in the data is because the probability of being ill is relatively low, and people choose not to take a physician office visit while the diease is not serious.

Question 5

```
p<-data.frame("Years"=c(12,12,32,20,20,27,23,19,23,26,21,3,8,35,2,19,8,25,33,35),"Salamanders"=c(3,4,8,
model.fit <- glm(Salamanders~Years, family="poisson", data=p)
summary(model.fit)
```

```
##
## Call:
## glm(formula = Salamanders ~ Years, family = "poisson", data = p)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0320  -0.8082  -0.1310   0.5307   2.2846
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.59136    0.29200   2.025   0.0428 *
## Years        0.04451    0.01136   3.919  8.9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 37.761  on 19  degrees of freedom
## Residual deviance: 21.219  on 18  degrees of freedom
## AIC: 88.648
##
## Number of Fisher Scoring iterations: 5
```

The Possion regression model is $Salamenders = e^{0.59136+0.04451*Yearsafterburn}$. The coefficient of years after burn is 0.0445, and $e^{0.0445} > 1$. The expected count $\mu = E(thenumberofsalamanders)$ is 1.045518 times larger than when years after burn is 0.

```
exp(coef(model.fit))
```

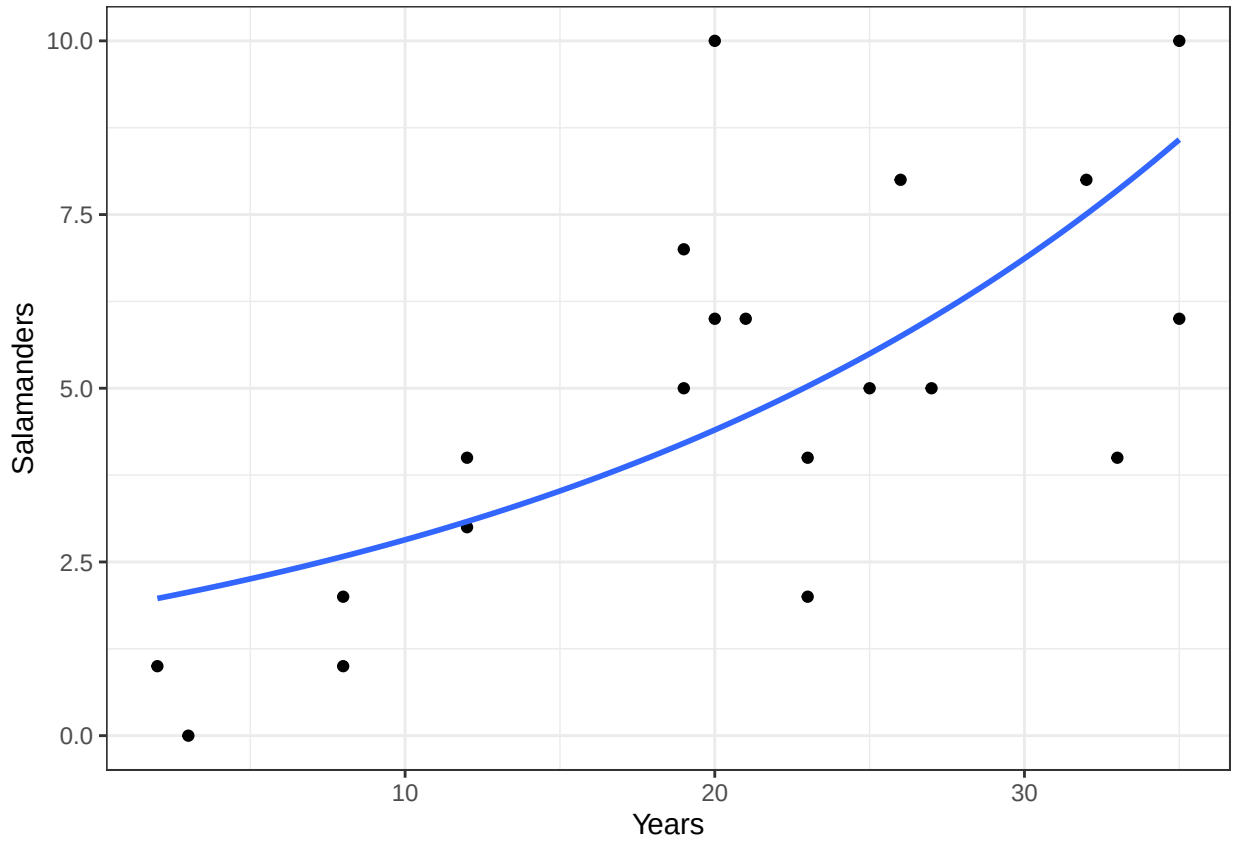
```
## (Intercept)       Years
##    1.806437    1.045518
```

```
library(ggplot2)
```

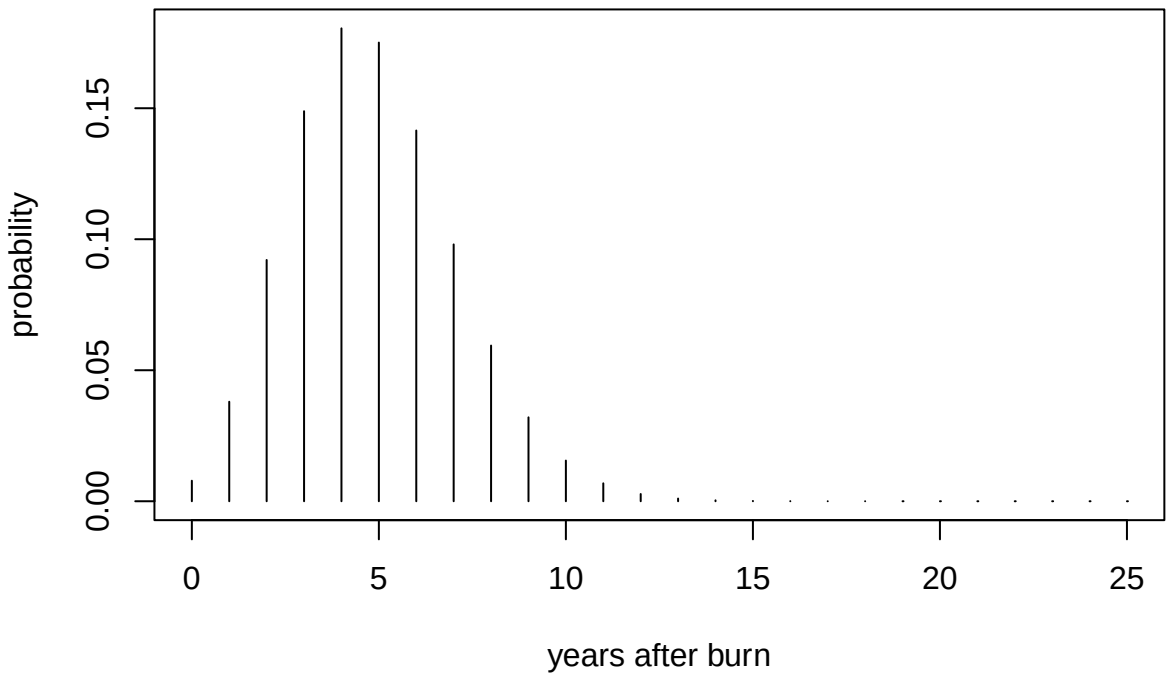
```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
ggplot(p, aes(Years, Salamanders)) +
  geom_point() +
  stat_smooth(method="glm",
              se=FALSE,
              method.args = list(family="poisson")) +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
years1<-c(12,12,32,20,20,27,23,19,23,26,21,3,8,35,2,19,8,25,33,35)
result<-exp(years1*0.04451+0.59136)
years2<-0:25
lambda=mean(result)
plot(years2, dpois(years2, lambda=lambda), type='h',xlab="years after burn",ylab = "probability")
```


probability
0.00
0.05
0.10
0.15
0
5
10
15
20
25
years after burn